SAPIENZA
UNIVERSITÀ DI ROMA

ISSN 2385-2755
DiSSE Working papers
[online]

WORKING PAPERS SERIES
DIPARTIMENTO DI
SCIENZE SOCIALI ED ECONOMICHE

Identification of Causal Mechanisms through an RD Approach

Viviana Celli

N. 15/2020

SAPIENZA - UNIVERSITY OF ROME
P.le Aldo Moro n.5 – 00185 Roma T(+39) 0649910563
CF80209930587 – P.IVA 02133771002

Identification of Causal Mechanisms through an RD Approach

Viviana Celli*

Abstract

Causal Mediation Analysis has important implications in economics. It helps to deeply understand the policy makers' decisions and to better design policy strategies. However, the identification process is not an easy issue and analyzing causal mechanisms requires stronger assumptions than evaluating the classical average treatment effect. The main difficulty consists in the endogeneity of the mediator with the consequence that it is not possible to identify the effects of interest. Several methods have been developed, based on different set of assumptions and with different strategies for the estimation. I propose a new identification strategy for the estimation of the direct and the indirect effect, through an implementation of a Regression Discontinuity Design. I present two different models. The first one follows the traditional identification strategy based on linear equation models. The second model follows the most recent literature based on non-parametric identification procedures. I show the consistency of this last estimator, validating the results through a Monte Carlo simulation study.

Keywords: Mediation Analysis, Regression Discontinuity Design, Direct effect, Indirect effect, Non-parametric identification
JEL Classification: C31, C54, D04

*Department of Social Sciences and Economics, Sapienza University of Rome. E-mail address: viviana.celli@uniroma1.it

1 Introduction

The vast majority of policy evaluation literature is concerned with the assessment of the causal effects of a public policy. More in general, the aim of causal analysis is to identify and estimate the effects of a treatment on the outcomes of interest, a parameter known in literature as Average Treatment Effect (ATE)[1]. Nevertheless, policy evaluation is not concerned about the channels through which the causal effect occurs. In the last decades, many studies have gone beyond the ATEs, focusing on the causal mechanisms through which the treatment transmits the effects to the outcome, a literature known as Causal Mediation Analiysis (see for instance Baron & Kenny, 1986; Pearl 2001; Imai et al. 2011). The idea of this kind of studies is to disentangle the total treatment effect into two components: the direct effect of the treatment on the outcome and the indirect effect, that operates through one or more intermediate variable, called mediators, that lie in the causal pathway between treatment and outcome. This analysis helps to understand the effects of the policymakers' decisions and to implement changes in the policy designs to make them more efficient with important and interesting implications. However, even in experimental designs, the causal mechanisms are not easy to identify. The main issue is that analyzing causal mechanisms requires stronger identifying assumptions than evaluating the classical ATE. First of all, the randomization of the treatment does not imply the randomness of the mediator, as discussed in Robins and Greenland (1992). The consequence is that the total effect cannot be disentangled by simply conditioning on the mediator, because this generally introduces selection bias coming from variables influencing both the mediator and the outcome (Rosembaum, 1984). Several methods have been developed, based on a different set of assumptions and with different strategies for the estimation. In particular, whereas earlier works often relied on tight linear specifications, as in Judd and Kenny (1981) and Baron and Kenny (1986), more recent studies focus on nonparametric and semiparametric identification, allowing for nonlinearities and heterogeneity in the effects of interest, as in Pearl (2001), Robins (2003), VanderWeele (2009), Imai, Keele and Yamamoto (2010) and Huber (2014), among many others.

The main contribution of this paper is to show the identification and the estimation of the natural direct and indirect effects with an implementation of a Regression Discontinuity Design (RDD) to solve the problem of the endogeneity of the mediator. Thanks to the presence of a continuous, observable forcing variable Z that generates variation in the state of the mediator, I can rule out the

[1] The parameters of interest of the causal analysis are also the Average Treatment Effect on Treated (ATET) and the Average Treatment Effect on Non Treated (ATENT).

presence of unobservable and observable post-treatment confounders that jointly affect the mediator and the outcome. In particular, I explain two different models, relying on two different sets of assumptions. The first one (Model 1) relies on a parametric identification and then is less flexible, recalling the linear structural equation model of an influential article by Baron and Kenny (1986). In particular in Model 1, the forcing variable Z is affected by the treatment and in turn deterministically affects the mediator.[2] The second model (Model 2) relies on a non-parametric identification and then it is based on a more flexible framework, in which it is not necessary the specification of the functional form of the variables and the heterogeneity of the effects is allowed (Imai et al., 2011). This is due to the fact that the forcing variable Z now is an exogenous variable but still deterministically affects the mediator.[3] In both cases I use a sharp RD, see for instance Trochim (1984), Imbens and Lemieux (2008), Lee (2008), Lee and Lemieux (2010), that permits to have compliers as population of reference. This is the first study that uses RD method to solve the problem of the endogeneity of the mediator and then it is the first methodological contribution that join these two literatures.[4] Moreover, there are very few studies using quasi-experimental designs inside the mediation framework[5]: an important contribution is given by E. Deuchert, M. Huber and M. Schelker (2019), who use a difference-in-differences approach for disentangling the total treatment effect, providing an empirical application based on the Vietnam draft lottery. Secondly, the estimation procedure is easy to implement, basing on a local regression to get the effect of the treatment on the outcome weighted by the treatment propensity scores that are straightforward to implement by a probit (or logit) estimation to get the potential values of the mediator. The estimation is computed in a bandwidth defined by $\bar{z}$.

The remainder of this paper is organized as follows. Section 2 defines the parameters of interest. Section 3 presents the model 1 and discusses the identifying assumptions, the parametric identification and gives a graphical interpretation. Section 4 presents the model 2 with its assumptions, the non-parametric identification and the graphical interpretation. Section 5 shows the estimation procedure and in section 6 I present a simulation study which shows the behavior of the estimators. Section 7 concludes.

[2] We could think about treatment like a job training program in which there are computer (PC) lectures. Z is a PC test score to measure the knowledge after the training and the mediator is a PC course that people have to attend only if they don't join the sufficiency in the test.

[3] Following the example in the previous footnote, the job training has not PC lectures, but the rule to attend the PC course is the same.

[4] An intuition about using RD to study indirect effects is given by Angelucci and Di Maro (2010)

[5] See Celli V. (2019) for a review of quasi-experimental designs in the Causal Mediation Analysis

2 Definition of parameters

The goal of Causal Mediation Analysis is to decompose the Average Treatment Effect (ATE) of a binary treatment D on the outcome Y into the indirect and the direct effect. The first one reflects one possible explanation for why treatment works, explicitly defining a particular mechanism behind the causal impact and it answers the following counterfactual question: what change would occur to the outcome if the mediator changed from what would be realized under the treatment condition, that is $M_i(1)$, to what would be observed under the control condition, that is $M_i(0)$, while holding the treatment status at d? The second one, the direct effect, represents all other possible explanations through which a treatment affects an outcome and it corresponds to the change in the potential outcome when exogenously varying the treatment but keeping the mediator fixed at its potential value $M_i(d)$. The estimation of these two effects is not an easy issue in empirical designs, because the endogeneity of the mediator implies the presence of post-treatment confounders, see for instance Imai et al. (2011). Even in the presence of a double randomization of the treatment and the mediator, we could still have the selection bias (Imai, Tingley and Yamamoto, 2013). To solve this problem I developed an identification strategy based on the counterfactual literature. In particular, I use a continuous and observed forcing variable Z[6] that can induce an exogenous change in the mediator state, depending if Z exceeds a known cutoff point z^*, recalling the Regression Discontinuity (RD) literature (David S. Lee, 2008; David S. Lee & T. Lemieux, 2010). Using the potential outcome framework permits to relax the assumptions behind the functional form of the variables and then to have a flexible identification procedure but, on the other hand, it implies to face the missing values problem (Holland, 1986), as I explain in the following sections.

2.1 Parameters of interest

To define the parameters of interest in this new setting that combines the mediation framework and the RD design I make use of potential outcome notation, see for instance Neyman (1923) and Rubin (1974). I denote by $Y(d', m)$ and $M(d)$ the potential outcome and the potential mediator state, with $d, d', m \in \{0, 1\}$. Furthermore, I denote by $Z{=}z^*$ the cutoff point at which the mediator state changes sharply, according to the following deterministic rule: $M_i{=}\{1[Z{\geq}\, z^*]\}$, where the subscripted i is the individual observation. I have two important implications thanks to this rule.

The first one is that, because of the status of M depends deterministically on

[6] In literature known also as *running* variable.

Z, there is no an error term in the selection into M, implying the absence of unobservable factors that could create the presence of Always takers, Never takers and Defiers in the behavior of M w.r.t. Z. In our sharp setting, we have only Compliers[7], meaning that who is above (or below[8]) z^* will have $M = 1$ ($M = 0$) and vice versa. In this way I can identify the potential value of the mediator, because I know for that population (Compliers) what would be the value of M under the opposite treatment status, simply looking at the control group.

The second key point is that, because I take only individuals just above and below the threshold, defined as $\bar{z} \in [z^* - \epsilon, z^* + \epsilon]$ according to the RD literature, the value of the mediator is like randomized in this window, meaning that units in the population of interest will have comparable observables and unobservables characteristics.

In this context I can locally define our parameters of interest as:

$$\theta(d) = E[1, M(d)) - Y(0, M(d))|Z = z^*], \qquad d \in \{0, 1\} \tag{1}$$

$\theta(d)$ is the average natural direct effect (Pearl, 2001)[9] for the population near the threshold and it expresses how much the mean potential outcome would change if the treatment was set from 1 to 0 but the mediator were kept at the potential level it would have taken in treatment status equal d. It captures what the effect of the treatment on the outcome would remain if we were to disable the pathway from the treatment to the mediator.

In the same way I can define the local natural average indirect effect as:

$$\delta(d) = E[Y(d, M(1)) - Y(d, M(0))|Z = z^*], \qquad d \in \{0, 1\} \tag{2}$$

$\delta(d)$ corresponds to the change in mean potential outcome for the population near the threshold when exogenously shifting the mediator to its potential values under treatment and non-treatment state but keeping the treatment fixed at $D = d$ to switch off the direct effect.

It can be easily shown that the ATE even for the local population is the sum of the natural direct and indirect effects defined upon opposite treatment states, like in the traditional mediation framework, but looking only at the individuals just above and below the cutoff point:

[7] The compliance is defined with respect to the forcing variable Z rather than the treatment.

[8] The interpretation depends on the definition of the score.

[9] Robins and Greenland (1992) and Robins (2003) denominated these parameters as "pure" direct and indirect effects.

$$\begin{aligned}\Delta &= E[(Y_1 - Y_0)|Z = z^*] \\ &= E[Y(1, M(1)) - Y(0, M(0))|Z = z^*] \\ &= E[Y(1, M(1)) - Y(0, M(1))|Z = z^*] + E[Y(0, M(1)) - Y(0, M(0))|Z = z^*] \\ &= [\theta(1) + \delta(0)|Z = z^*] \\ &= E[Y(1, M(0)) - Y(0, M(0))|Z = z^*] + E[Y(1, M(1)) - Y(1, M(0))|Z = z^*] \\ &= [\theta(0) + \delta(1)|Z = z^*]\end{aligned} \tag{3}$$

where the third equality comes from adding and subtracting the quantity $E[Y(0, M(1))]$ and the fifth equality comes from adding and subtracting the quantity $E[Y(1, M(0))]$. The main problem with this analysis is identifying the counterfactual quantities $E[\,Y(d,\ M(d'))]$, never observed for each individual and hardly identified in non-experimental designs with the classical assumptions. A second issue is that only one of $Y(1, M(1))$ and $Y(0, M(0))$ is observed for any unit, which is known as the fundamental problem of causal inference (Holland, 1986).
Identification of direct and indirect effect hinges on exploiting exogenous variation in the treatment and the mediator, as follows in the next section.

2.2 Natural and controlled effects

Another parameter taken into account from the mediation literature is the controlled direct effect (CDE). Formally:

$$\gamma(m)^\star = E[Y(1, m) - Y(0, m)], \qquad \text{for } m \text{ in the support of } M \tag{4}$$

and it expresses how much the mean potential outcome would change if the mediator were fixed at a particular value m uniformly in the population but the treatment was exogeneously changed from 1 to 0. Usually it is easier identify this parameter because it is not necessary to know the potential value of the mediator and then the analyst needs less assumptions. At the same time, for the policy implications most of the time is useful to know the natural effects. Unfortunately, the CDE is equal to NDE only if there is no interaction effect between treatment and mediator to the outcome and then in the absence of heterogeneity.
But, as before, I menage a local controlled direct effect, defined just for the population near the threshold:

$$\gamma(m) = E[Y(1, m) - Y(0, m)|Z = z^*], \qquad \text{for } m \text{ in the support of } M \tag{5}$$

The main implication of this local analysis is that, because we are in a sharp RD, meaning that M is deterministically determined by Z, everyone is complier in our population of interest, implying that $[\gamma(m{=}0)|Z{=}z^*] = [\theta(0)|Z{=}z^*]$. In fact, the direct effects reflect the difference in the outcomes between treated and non-treated groups, maintaining a fixed level of M. But, at the threshold, I know that the entire population under analysis is complier, so everyone who has $Z \leq z^*$ has $M{=}0$ and this permits to identify not only $Y(1,\ 0)$ but also $Y(1,\ M(0))$, because I know for everyone what would be the potential mediator defined on the opposite status of the treatment. If in reality the behavior is like in the threshold I can identify the natural effects because, in this setting, the CDE coincides with the NDE. It would no longer be true if we were in a Fuzzy RD context, because we should control for the presence of Never Takers, Always takers and Defiers. In this case the CDE is no longer equal to the NDE, because the reference population will be different, and we couldn't know the potential value of the mediator simply looking at the value of Z.
In our framework the parameters of interest are:

$$NDE(d) = E[Y(1, M(d)) - Y(0, M(d))|Z = z^*] \quad \forall\ d \in \{0, 1\} \tag{6}$$

$$NIE(d) = E[Y(d, M(1)) - Y(d, M(0))|Z = z^*] \quad \forall\ d \in \{0, 1\} \tag{7}$$

In the next sections, I will discuss two different models.

3 Model 1

I consider a first general model in which a random binary treatment D affects the outcome Y and the forcing variable Z. This one deterministically affects the mediator M, inducing a sharp change in the mediator state depending on the particular value of Z, and in turn it affects the outcome Y. In this model a causal effect between Z and Y is allowed, because to estimate the effects of interest we have to look just at the population near the threshold $Z{=}\bar{z}$, controlling, then, for the direct effect of the forcing variable on the outcome. So, in this model, Z is a continuous, observed and endogenous variable.
The general model is given by:

$$\begin{gathered} Y = \phi(D, Z, M, u) \\ M = 1[Z \geq z^*] \\ Z = \xi(D, v) \\ D = \lambda(\epsilon) \end{gathered}$$

where ϕ, ξ, λ are linear functions and u, v, ϵ are unobservable components. In the model's notation I didn't include the set of covariates X for sake of simplicity. The general outcome equation is:

$$Y = \phi\{D(\epsilon), Z[D(\epsilon), v], M[Z(D(\epsilon), v)], u\}$$

3.1 Identifying assumptions of Model 1

For the first model I assume that the forcing variable Z is function of the treatment D. To identify our parameters of interest the first assumption I need is the classical conditional independence of the treatment, see for instance Imbens (2004):

ASSUMPTION 1. Conditional randomness of the treatment:

$$\{Y(d', m), M(d)\} \perp D|X = x, \qquad \forall\ d, d', m \in \{0, 1\}$$

By assumption 1 I state that there are no unobserved confounders between treatment and mediator and/or outcome conditioning on pre-treatment covariates X, implying the independence of the potential outcome and the potential mediator from D. With this assumption I can identify the direct effect from D to Y and the effect from D to M. In non-experimental data, the plausibility of this assumption depends on the richness of variables available. In experimental data, this assumption holds if the treatment is randomized within strata defined on X.[10]

ASSUMPTION 2. Continuity of the potential outcome at the threshold:

$$E\{Y(d', m)|Z, X\} \quad \text{is continuous in } Z = z^*$$

This assumption states that in the potential outcomes there is no discontinuity due to selection bias and that conditioning on Z and X, M is like randomized at the threshold, implying then the absence of unobserved confounders jointly affecting the mediator and the outcome, recalling the RD literature's assumptions (see Lee, 2008). In other words, the jump observed in the factual outcome at the threshold is only due to the mediator. It means that near the threshold I can correctly identify the causal effect of M on Y. The difference with the classical mediation framework is that now I have to look at a local population. Looking at the threshold also permits to do not take into account the relation between Z

[10] If treatment is randomized unconditionally, the stronger assumption $\{Y(d', m), M(d)\} \perp D$ holds as well.

and Y, implying an exlusion restriction for Z, such that $corr(Z, Y) = 0|Z = z^*$. In this way the indirect effect is not confounded by Z for the local population. Assumption 2 is violated if unobserved pre-treatment confounders affect both M and Y directly, or if unobserved post-treatment variables influence M and Y and are not fully determined by X and/or D.

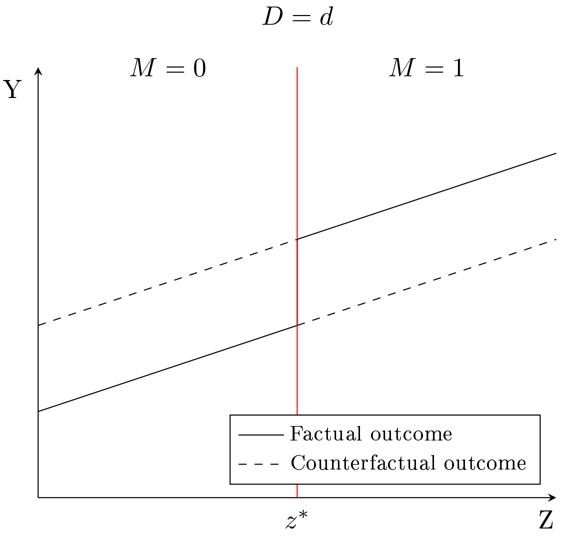

Figura 1: Assumption 2

ASSUMPTION 3. Perfect compliance of the mediator:

$$P(M = 1|Z = z^*_+) = 1$$
$$P(M = 0|Z = z^*_-) = 1$$

This assumption comes from the Sharp RD design (SRDD) and it states that the mediator is deterministically and fully determined by the value of Z, meaning that the effect of D goes to M only through Z, implying an exclusion restriction w.r.t. the treatment. In particular, every observation with a score just above z^* will have M=1 and every observation with a score just below z^* will have M=0.[11] It is important to note that there is no error term, implying the presence only of Compliers in our population of interest. In other words, for this population, I know for sure what will be the value of the mediator, simply

[11] This deterministic law can be written as: $M_i = 1[Z \geq z^*]$.

looking at the score of Z.

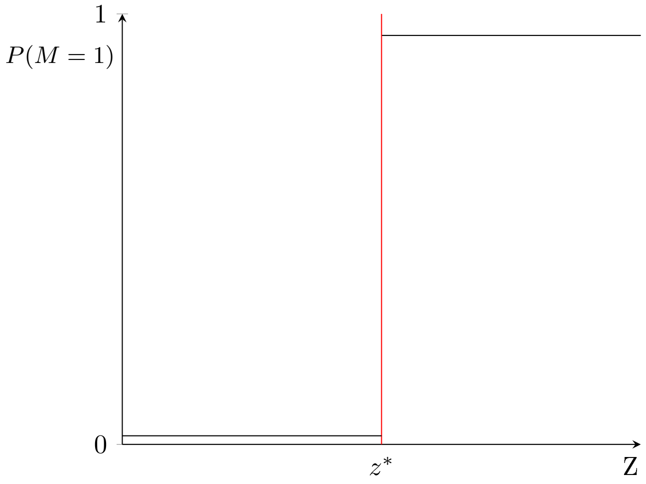

Figura 2: Assumption 3

ASSUMPTION 4. Homogeneity effect:
I can add a parametric assumption of effect homogeneity:

$$\theta(0) = \theta(1) = \theta$$

$$\delta(0) = \delta(1) = \delta$$

This assumption states that direct and indirect effects do not vary as functions of treatment status. In other words, the direct and indirect effects are independent of each other and the direct effect is constant regardless the level of the mediator and the indirect effect is constant regardless the level of the treatment removing the possibility of non-linearity, see for instance VanderWeele (2015). This assumption is necessary for the identification strategy, as I explain later. Assumptions 1-4 imply:

ASSUMPTION 5. Conditional indipendence between treatment and forcing variable Z:

$$Z(d) \perp D|X = x, \qquad \forall\ d \in \{0, 1\}$$

This assumption holds thanks to assumption 1 and 4. Without the parametric assumption this would no longer be true. This assumption states that if there are no confounders between treatment and mediator as stated by Assumption 1, then there are no confounders between treatment and Z if at the threshold the entire effect of D goes to M through Z. This implies that I can correctly identify the homogeneous effect from D to Z.

3.2 Parametric identification of Model 1

By Assumptions 1-5, I can parametrically identify model 1. According to the literature, the parametric (linear) estimation implies the absence of interactions between the effect of D and M on Y and it imposes additivity between the observed and unobserved terms, implying that the effects are constant across individual characteristics.[12] So, by Assumption 4, I can rewrite a parametric (but less flexible) model like:

$$\begin{cases} Y = \beta_0 + \beta_1 D + \beta_2 M + \beta_3 Z + u \\ M = 1[Z \geq \bar{z}] \\ Z = \alpha_0 + \alpha_1 D + v \end{cases}$$

This system recalls the linear structural equation model (LSEM), see for instance Baron & Kenny (1984).
Assuming $\beta_0 = \alpha_0 = 0$ for the sake of simplicity, I can rewrite the linear outcome equation as:

$$\begin{gathered} Y = D(\beta_1 + \beta_3\alpha_1) + \beta_2 M + (\beta_2 v + u) \\ with \quad M = 1[Z \geq \bar{z}] \end{gathered} \tag{8}$$

where, in (8), β_1 represents the direct effect of D on Y, $\beta_3\alpha_1$ is the partial indirect effect that goes from D to Y through Z and β_2 is the part of the indirect effect that goes from M to Y if $Z \geq z^*$, otherwise the indirect effect due to the mediator will be null. It is worth to note that, in this setup, the total effect is given simply by summing the direct and indirect effects. In the classical policy analysis we observe only one coefficient for D that can be unbiased if correctly specified but it can't explain the "causes of the effect" but only the "effects of the causes".
I can identify these effects because by Assumption 1 the direct effect is unconfounded; always by Assumption 1 and 5 I can still correctly estimate the effect from D to Z; by Assumption 3 I know that M is deterministically determined by

[12] The model can be augmented adding the interaction term between mediator and treatment. This, at least, allows for an heterogeneity in $\theta(d)$ and $\delta(d)$ w.r.t. d

Z, implying unconfoundedness of M; even if the corr(u,v)$\neq$0 it doesn't matter because I have to look at the threshold by Assumption 2; then, because we are in a LSEM I can estimate the indirect effect simply multiplying every single causal parameter. I can't obtain a non-parametric identification because without the homogeneity assumption I am not able to correctly estimate the effects, because of the correlation between D and v when I condition on a particular state of treatment and on the value of Z. Because of this correlation Assumption 5 does not hold and the non-parametric identification is impossible to join.

3.3 Graphical interpretation of Model 1

The first Model can be represented like in Figure 3: it illustrates the framework based on a direct acyclic graph, in which the arrows represent causal observable effects and the dashed arrows represent unobservable effects. I didn't take into account the set of covariates X for ease of exposition.

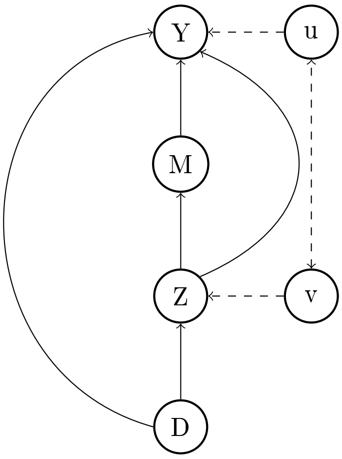

Figure 3: Model 1

This causal diagram satisfies Assumotions 1-5. In particular, it shows the effects that can be parametrically identified for the local population. An important point is that actually Z is a post-confounder variable because it is caused by the treatment and in turn it causally affects both M and Y. But, because I am interested only on the local population, fixing $Z = \bar{z}$ I can correctly estimate the causal effect that goes from D to Y through Z and M, allowing also for a correlation between the error terms of the forcing variable and the outcome, respectively u and v. The limit of this setup is that fixing both the treatment and the forcing variable I can't nonparametrically identify the effects, because even if I have had the randomization of the treatment, I would have had selection bias in the estimation of the effects.

4 Model 2

I consider a second general model in which the treatment D affects the outcome and the mediator, but now the forcing variable Z[13] is not affected by treatment, but still deterministically affects the mediator, like in Figure 4 . In particular, D is randomly assigned and Z is an exogenous variable, implying a zero correlation between the error terms of these two variables. So, the mediator is a function of the treatment state and it sharply changes state depending on the value of Z. The general model is given by:

$$\begin{aligned} Y &= \phi(D, M, Z, u) \\ M &= \zeta(D, Z) \\ Z &= \xi(v) \\ D &= \lambda(\epsilon) \end{aligned}$$

where ϕ, ζ, ξ, λ are unknown functions and u, v, ϵ are unobservable components. I didn't include in the model's notation the set of covariates X for ease of exposition, but the assumptions discussed later on are more plausible after conditioning on observable characteristics.
The general outcome equation is given by:

$$Y = \phi[D(\epsilon), Z(v), M(D(\epsilon), Z(v), u)]$$

4.1 Identifying assumptions of Model 2

If we are in a framework in which Z is an exogenous forcing variable I need a different set of assumptions for the identification of the effects of interest. In particular, in addition to assumptions 1 and 2 I have:

ASSUMPTION 7. Conditional indipendence between the treatment and the forcing variable:

$$Z \perp D | X = x$$

meaning that now Z is not a function of the treatment and it is still orthogonal to the treatment conditional on X. But now, this assumption holds even without the homogeneity assumption required in Model 1[14], because I don't have correlation between v and ϵ. This means that I can identify the effects even if

[13] The forcing variable Z must be always a continuous and observable variable.

[14] See Assumption 4

they are not constant across units, allowing for a more flexible design.
Assumptions 1 and 7 imply:

ASSUMPTION 8. Conditional randomness of the treatment at the threshold:

$$\{Y(d', m), M(d)\} \perp D|Z = z^*, X = x \qquad \forall\, d, d', m \in \{0, 1\}$$

This assumption is implied by Assumption 7. Under the absence of homogeneity, the effects that are identified under the continuity assumption, are local effects that are specific to the population with $Z = \bar{z}$. It permits to have a non-parametric identification of the natural effects, because now I don't have correlation between treatment and the error term of Z and then I have the independence between the treatment and the potential outcome at the threshold. Assumption 8 is weaker than assumption 1, because now the treatment can be randomized only at the threshold.

ASSUMPTION 9. Compliance of the mediator:

$$Pr(M = 1|Z^+, D = 1) = 1$$
$$Pr(M = 0|Z^-, D = 1) = 1$$
$$Pr(M = 0|D = 0) = 1$$

In this model the mediator is a deterministic function of D and Z. In particular, in the treated group I can observe two different values of M depending on the cutoff point z^*. On the contrary, in the control group I observe just the mediator status equal zero. This implies that I cannot identify all the parameters of interest, but still I can identify some effects under analysis.

ASSUMPTION 10. Common support:
$0 < Pr(D = d|Z = \bar{z}, X = x) < 1, \qquad \forall\, d \in \{0, 1\}$ and x in the support of X

By Assumption 10, the conditional probability to receive or not receive the treatment given Z and X is between 0 and 1, meaning that I can observe a particular value of Z and X both in the treated and non treated group. This assumption is stronger than the standard common support in policy evaluation.[15] By Bayes' theorem, Assumption 10 also implies that $0 < Pr(Z = \bar{z}|D = d, X = x) < 1$, meaning that conditional on X, the forcing variable must not be a deterministic function of the treatment, otherwise no comparable units would be available across treatment states.

[15] $0 < Pr(D = d|X = x) < 1$, for each value of X there is a positive probability of being both treated and untreated.

4.2 Non-Parametric identification of Model 2

Now, Assumption 1 and Assumption 7 imply Assumption 8 and this allows for a local non-parametric (and more flexible) identification of the natural effects. In particular, I can identify the counterfactual quantity $E[Y(d, M(d'))]$:

$$\begin{aligned}
&E[Y(d,M(d'))|Z=\bar{z}] = \\
=&\iint E[Y(d,m)|M(d')=m, Z=\bar{z}, X=x]\, dF_{M(d')|Z=\bar{z},X=x}(m)\, dF_{x|Z=\bar{z}}(x) \\
=&\iint E[Y(d,m)|M(d')=m, D=d', Z=\bar{z}, X=x]\, dF_{M(d')|Z=\bar{z},X=x}(m)\, dF_{x|Z=\bar{z}}(x) \\
=&\iint E[Y(d,m)|D=d', Z=\bar{z}, X=x]\, dF_{M(d')|Z=\bar{z},X=x}(m)\, dF_{x|Z=\bar{z}}(x) \\
=&\iint E[Y(d,m)|D=d, Z=\bar{z}, X=x]\, dF_{M(d')|D=d',Z=\bar{z},X=x}(m)\, dF_{x|Z=\bar{z}}(x) \\
=&\iint E[Y(d,m)|M=m, D=d, Z=\bar{z}, X=x]\, dF_{M|D=d',Z=\bar{z},X=x}(m)\, dF_{x|Z=\bar{z}}(x) \\
=&\iint E[Y|M=m, D=d, Z=\bar{z}, X=x]\, dF_{M|D=d',Z=\bar{z},X=x}(m)\, dF_{x|Z=\bar{z}}(x)
\end{aligned} \tag{9}$$

$$\begin{aligned}
=&\iint E[Y|M=m, D=d, Z=\bar{z}, X=x] \cdot \frac{Pr(D=d'|M=m, Z=\bar{z}, X=x)}{Pr(D=d'|Z=\bar{z}, X=x)} \\
&dF_{M|Z=\bar{z},X=x}(m)\, dF_{X|Z=\bar{z}}(x) \\
=&\int E[Y|M=m, D=d, Z=\bar{z}, X=x] \cdot \frac{Pr(D=d'|M=m, Z=\bar{z}, X=x)}{Pr(D=d'|Z=\bar{z}, X=x)}\, dF_{M=m,X=x|Z=\bar{z}}(m,x) \\
=&E\left[E\Big[Y|M=m, D=d, Z=\bar{z}, X=x\Big] \cdot \frac{Pr(D=d'|M=m, Z=\bar{z}, X=x)}{Pr(D=d'|Z=\bar{z}, X=x)} \middle| Z=\bar{z} \right]
\end{aligned} \tag{10}$$

The first equality follows from the law of iterated expectations and from replacing the outer expectations by integrals, the second from Assumption 1 and 7, the third from Assumption 2, the fourth from Assumtion 1 and 7 again, the fifth from Assumption 2, the sixth from Assumption 1, the seventh and the eighth equality follows from Bayes' theorem and the last one from the law of iterated expectations.

(9) recalls the so called mediation formula for identifying the direct and indirect effect, see for instance Pearl (2001) and Imai, Keele and Yamamoto (2010), with the difference that now we are looking only at the population near the threshold.

With weaker restrictions, I can identify the observable quantity $E[Y(d, M(d))|Z = \bar{z}]$:

$$\begin{aligned} &E[Y(d,M(d))|Z=\bar{z}] = \\ =&E\Big[E[Y(d,M(d))|Z=\bar{z},X=x]\Big|Z=\bar{z}\Big] \\ =&E\Big[E[Y|D=d,Z=\bar{z},X=x]\Big|Z=\bar{z}\Big] \end{aligned} \tag{11}$$

where the first equality follows from the law of iterated expectation and the second from Assumption 1 and 7.

Therefore, $\theta(d)$ and $\delta(d)$ are identified by either subtracting (10) from the equation (11) or vice versa, depending on whether d is one or zero. In particular, the average direct effect $\theta(d)$ is given by:

$$\begin{aligned} \theta(d) =& \iint \Big[E[Y|D=d,M=m,Z=\bar{z},X=x] - E[Y|D=d',M=m,Z=\bar{z},X=x]\Big] \\ & dF_{M|D=d,Z=\bar{z},X=x}(m)\, dF_{x|Z=\bar{z}}(x) \\ =& \iint \Big[E[Y|D=d,M=m,Z=\bar{z},X=x] - E[Y|D=d',M=m,Z=\bar{z},X=x]\Big] \\ & \cdot \frac{Pr(D=d|M=m,Z=\bar{z},X=x)}{Pr(D=d|X=x,Z=\bar{z})}\, dF_{M=m,X=x|Z=\bar{z}}(m,x) \\ =& E\Bigg[\Big[E[Y|D=d,M=m,Z=\bar{z},X=x] - E[Y|D=d',M=m,Z=\bar{z},X=x]\Big] \\ & \cdot \frac{Pr(D=d|M=m,Z=\bar{z},X=x)}{Pr(D=d|Z=\bar{z},X=x)}\Bigg|Z=\bar{z}\Bigg] \end{aligned} \tag{12}$$

whereas the indirect effect $\delta(d)$ is given by:

$$\begin{aligned} \delta(d) =& \iint E[Y|D=d,M=m,Z=\bar{z},X=x] \cdot \\ & \{\, dF_{M=m|D=d,Z=\bar{z},X=x}(m) - \, dF_{M=m|D=d',Z=\bar{z},X=x}(m)\}\, dF_{x|Z=\bar{z}}(x) \\ =& E\Bigg[E\Big[Y|D=d,M=m,Z=\bar{z},X=x\Big] \cdot \\ & \left(\frac{Pr(D=d|M=m,Z=\bar{z},X=x)}{Pr(D=d|Z=\bar{z},X=x)} - \frac{Pr(D=d'|M=m,Z=\bar{z},X=x)}{Pr(D=d'|Z=\bar{z},X=x)}\right)\Bigg|Z=\bar{z}\Bigg] \end{aligned} \tag{13}$$

Following the identification results and assuming the availability of an i.i.d. sample of size n, I can estimate the natural direct effect under control group $\theta(0)$ and the natural indirect effect under treated group $\delta(1)$ and the total effect given by the sum of the previous two effects[16]. In general, they can be estimated by various strategies. In literature, parametric methods have been commonly used, like in Pearl (2011) and VanderWeele(2009), but they have some drawbacks like a restrictive functional form and a difficult interpretability in case of nonlinearities. Most recent nonparametric estimation has been developed by Imai et al. (2010). These methods avoid the before mentioned shortcomings, but they might be cumbersome in an empirical application in case of high dimensionality of X or if M is continuous. In this case the estimation is based on a combination of them. In particular, I estimate the conditional mean of Y by local regression and by a weighting formula the density of M, once I conditioned on a particular window defined in Z, recalling the RD estimation strategy.

4.3 Graphical interpretation of Model 2

Model 2 can be represented by the following acyclical graph of causal relations between observed and unobserved variables, in which for sake of simplicity I neglected the set of covariates X:

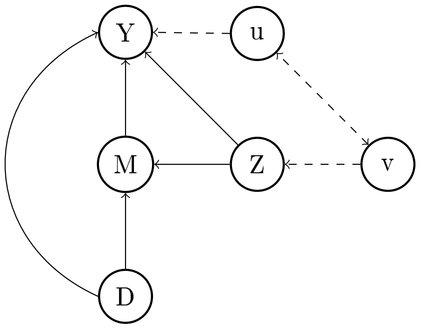

Figure 4: Model 2

The assumptions before discussed satisfy our structural system represented in Figure 4. In particular, by assumption 1 and 7 I can correctly identify the effect of the treatment on the mediator and by assumption 8 I can control for the direct effect from Z to Y, recalling the exclusion restriction for the forcing variable. In other words, I can identify the natural effects, even in the presence of a correlation between u and v. On the contrary, in this model, the correlation

[16] By Assumption 9, it is not possible to identify $\theta(1)$ and $\delta(0)$.

between v and the error term of the treatment is not allowed, as stated by assumption 7.

5 Estimation

Estimation based on the mediation formula (9) requires plug-in estimates for the conditional mean outcomes and the conditional mediator densities. In our model, the estimators of direct and indirect effects are given by:

$$\hat{\theta}(0) = \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\hat{\mu}_Y(1, M_i, Z_i, X_i) - \hat{\mu}_Y(0, M_i, Z_i, X_i)\right]\left(\frac{\hat{\rho}(m_i, x_i)}{1 - \hat{p}(x_i)}\right)\middle| Z = \bar{z}\right\} \tag{14}$$

$$\hat{\delta}(1) = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mu}_Y(1, M_i, Z_i, X_i)\left(\frac{\hat{\rho}(m_i, x_i)}{\hat{p}(x_i)} - \frac{1 - \hat{\rho}(m_i, x_i)}{1 - \hat{p}(x_i)}\right)\middle| Z = \bar{z}\right\} \tag{15}$$

where I define the bandwidth by a local linear regressions performed to either side of the cutpoint using the Imbens-Kalyanamaran optimal bandwidth calculation (2012) and $\hat{\rho}(m_i, x_i)$ and $1 - \hat{p}(x_i)$ denote the respective estimates of the propensity scores $Pr(D = 1|M = m_i, X = x_i)$ and $Pr(D = 1|X = x_i)$. Treatment propensity scores might be estimated by probit or logit specifications, see for instance Huber (2014) and Tchetgen Tchetgen (2013). The model can be better specified adding the interaction term between all variables' combinations in the conditional means outcome.

6 Simulation Study

This section presents a simulation study that provides some intuition for the identification result. I consider a data generating process (DGP) based on the following equations:

$$Y = 5 \cdot D + 3 \cdot M + 0.5 \cdot Z + \beta_1 \cdot DZ + \beta_2 \cdot ZM + 2 \cdot X + \epsilon_Y \tag{16}$$

$$M = I(D \cdot Z > 0) \tag{17}$$

$$D = I(0.5 \cdot x + 2 \cdot \epsilon_D > 0) \quad \text{with} \quad \epsilon_D, \epsilon_Y \sim N(0,1) \quad \textit{i.i.d.} \tag{18}$$

Equation (16) is the outcome equation, in which the observed Y is function of the observed variables D, M, Z, X and of the unobserved term ϵ_Y. β_1 and β_2

capture the interaction between respectively D and Z and Z and M. Equation (17) describes the mediator behavior under Assumption 9. In the simulations, I set $\beta_1 = 1.3$ and $\beta_2 = 1.5$[17]. Table I provides the true direct and indirect effects.

Table I. True effects

$\theta(0)$	5
$\delta(1)$	1.5
Δ	6.5

I run two simulation studies with a bandwidth chosen using the Imbens-Kalyanaraman optimal bandwidth calculation[18]. In the first scenario I have 1000 observations, whereas in the second scenario I have 2000 observations.

Table II and III present the bias, variance (VAR) and root mean squared error (RMSE) of the estimators in the two scenarios. As the tables show, augmenting the number of simulations $\hat{\theta}(0)$ performs much better, reaching zero bias. The behavior of $\hat{\delta}(1)$ is slower, but it respects the asymptotic characteristics. Moreover, in each simulation I applied the trimming rule (with trim=0.05) to discard observations with extreme propensity scores to improve overlap. The default is to discard observations with treatment propensity score smaller than 0.05 (5%) or larger than 0.95 (95%).

Table II

	Δ	$\theta(0)$	$\theta(0)$ trim	$\delta(1)$	$\delta(1)$ trim
Bias	-0.041	0,0016	0,0016	-0.043	-0.043
VAR	0,088	0,111	0,111	0,0871	0,0871
RMSE	0,096	0,111	0,111	0,095	0,095
nsim = 1000; nobs = 1000					

Table III

	Δ	$\theta(0)$	$\theta(0)$ trim	$\delta(1)$	$\delta(1)$ trim
Bias	0,024	0,000	0,0001	0,024	0,024
VAR	0,09	0,0637	0,0637	0,03	0,03
RMSE	0,08	0,06	0,06	0,037	0,037
nsim = 2000; nobs = 1000					

[17] I chose these values following other simulation studies in the mediation literature.

[18] See the RDestimate R package.

7 Conclusions

This paper presents a new estimator to identify average natural direct and indirect effects of a binary treatment using an RD approach. Identification relies on the local analysis of the effects, defined within a threshold $\bar{z}$, determined by the forcing variable Z. I have considered two different models and, then, two sets of assumptions: (i) in the first model, the forcing variable Z is causally affected by the treatment and in turn deterministically affects the mediator and (ii) in the second model, Z is no more influenced by the treatment but still deterministically influences the mediator. It has been shown that the effects can be identified in either case, but in the first model only a parametric identification is feasible, whereas in the second model a nonparametric identification is possible, allowing for a more flexible and realistic model's interpretation.

References

Angelucci M., Di Maro V. (2015): "Program evaluation and spillover effects", *Working paper, University of Michigan.*

Baron R.M., Kenny D.A. (1986): "The moderator-mediator variable distiction in social psychological research: conceptual, strategic and statistical considerations", *Journal of Personality and Social Psychology*, 51, 1173-1182.

Brader T., N. A. Valentino and E. Suhay (2008): "What triggers public opposition to immigration? Anxiey, group cues and immigration", *American Journal of Political Sciences*, 52(4), 959-978.

Celli, V. (2019): "Causal Mediation Analysis in Economics: objectives, assumptions, models" *Mimeo*, Sapienza University of Rome.

Deuchert, E., M. Huber, M. Schelker (2019): "Direct and indirect effects based on difference-in-differences with an application to political preferences following the Vietnam draft lottery", *Journal of Business & Economic Statistics*, 37(4), 710-720.

Flores C.A., Flores-Lagunes A. (2009): "Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness", IZA Discussion Paper no. 4237.

Frölich M., Huber M. (2017): "Direct and indirect treatment effects causal chains and mediation analysis with instrumental variables", *Journal of the Royal Statistical Society: Serie B*, 79(5), 1645-1666.

Holland, P. W. (1986): "Statistics and causal inference", *Journal of the American Statistical Association*, 81(396), 945-960.

Huber M. (2014): "Identifying causal mechanisms (primarly) based on inverse probability weighting", *Journal of Applied Econometrics*, 29, 920-943.

Imai K., Keele L., Yamamoto T. (2010): "Identification, inference and sensitivity analysis for causal mediation effects", *Statistical Science*, 25(1), 51-71.

Imai K., Tingley D., Yamamoto T. (2013): "Experimental designs for identifying causal mechanisms", *Journal of the Royal Statistical Society, Series A*, 176(1), 5-51.

Imai K., Keele L., Tingley D., Yamamoto T. (2011): "Unpacking the black box: learning about causal mechanisms from experimental and observational studies", *Political Science Review*, 105, 765-789.

Imbens G.W. (2004): "Nonparametric estimation of average treatment effects under exogeneity: a review", *The Review of Economics and Statistics*, 86(1), 4-29.

Imbens G.W., Kalyanamaran K. (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator", *The Review of Economic Studies*, 79(3), 933-959.

Imbens G.W., Lemieux T. (2008): "Regression discontinuity designs: a guide to practice", *Journal of Econometrics*, 142(2), 615-635.

Judd C.M., Kenny D.A. (1981): "Process analysis: estimating mediation in treatment evaluations", *Evaluation Review*, 5(5), 602-619.

Lee, D.S. (2008): "Randomized experiments from non-random selection in U.S. House Election", *Journal of Econometrics*, 142(2), 675-697.

Lee, D.S., Lemieux T. (2010): "Regression discontinuity designs in economics", *Journal of Economic Literature*, 48(2), 281-355.

Neyman, J. (1923): "On the application of probability theory to agricultural experiments: essay on principles, section 9", *Translated in Statistical Science*, 5, 465-480, 1990.

Pearl J. (2001): "Direct and indirect effects. In proceedings of the 17th Conference on uncertainty in artificial intelligence, Morgan Kaufman: San Francisco, 411-420.

Pearl J. (2011): "The causal mediation formula: a practitioner guide to assessment of causal pathways. Technical Report R-379, University of California, LA.

Robins J.M. (2003): "Semantics of causal DAG models and the identification of direct and indirect effects. In highly structured stochastic system, Green P., Hjort N., Richardson S., *Oxford University Press*, Oxford, 70-81.

Robins J.M., Greenland S. (1992): "Identifiability and exchangeability for direct and indirect effect", *Epidemiology*, 3(2), 143-155.

Rosembaum P. (1984): "The consequences of adjustment for a concomitant variable that has been affected by the treatment", *Journal of Royal Statistical Society, Series A*, 147(5), 656-666.

Rubin D.B. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies", *Jourbal of Educational Psychology*, 66(5), 688-701.

Tchetgen Tchetgen, E. J. (2013): "Inverse odds ratio-weighted estimation for causal mediation analysis", *Statistics in Medicine*, 32(26), 4567-4580.

Trochim W. M. K. (1984): "Research design for program evaluation: the regression-discontinuity approach", Sage publications, Beverly Hills, CA.

VanderWeele T.J. (2009): "Marginal structural models for the estimation of direct and indirect effects", *Epidemiology*, 20(1), 18-26.

VanderWeele T.J. (2015): "Explanation in causal inference. Methods for mediation and interaction", Oxford University Press.